경시적자료분석 기말 보고서

이름 : 정희철

논문 : Bayesian Joint Modelling of Longitudinal and Time-to-Event Data

저자 : Maha Alsefri

선택한 논문의 제목은 “Bayesian Joint Modelling of Longitudinal and Time-to-Event Data”이다. 해당 논문은 바이오 분야에서 많이 쓰이는 Joint Modelling에 대해 다양한 소개를 해주는데, 이 때 Bayesian으로 정한 이유는 이미 Frequentist 방법론들은 많은 연구가 진행되었고, 논문이 나와있기 때문이라고 하였다. Joint Modelling이란, 경시적 자료Longitudinal Data와 생존 자료 Time-to-Event Data의 관계성을 이용하여 더욱 정확한 설명력을 가지는 것을 목표로 하는 분석 방법이다. 이 때, 생존 자료Time-to-Event Data린 어떠한 사건이 일어났는가 와 해당 사건이 언제 일어났는지에 대한 정보가 담긴 자료로, 생존분석이 중용되는 바이오 분야에서 활발히 사용되고 있다.

저자는 관련 논문들을 모두 읽은 뒤 논문 별로 Response Variable이 일변량 또는 다변량인지, 그리고 데이터 유형이 연속형인지, Count인지, 어떤 모델링 기법을 사용했는지, 어떤 Error Distribution을 가정했는지, 어떤 Association Structure을 사용했는지 등등 매우 상세부분으로 나눠 각 방법 별 비율을 제시하였다.

저자가 논문을 모으고 분류한 방법은 아래 그림 1을 참고하면서 설명하겠다. Medline, Scopus, Web of Science에서 “joint model AND Bayesian”, “joint models AND Bayesian”, “joint modelling AND Bayesian”, “longitudinal AND Bayesian”, “survival AND Bayesian” 키워드들을 사용하여 나오는 논문 각 179개, 412개, 206개의 논문들을 찾을 수 있었고, 이 중 중복된 논문들을 제거하여 495개를 수집하였다. 이 중, 하나씩 읽어보면서 나열된 기준들을 만족하지 못하면 폐기하여 최종적으로 75개의 methodological 논문들이 남았고, 이 논문들을 사용하여 앞서 기술한대로 세부사항들로 분류하였다.

Diagram

Description automatically generated

그림 1 : 논문 수집 및 분류 과정

각 세부사항마다 상당히 많은 기법 또는 가정이나 이론들이 소개되어 있었다. 논문 내용을 충실히 반영하게 된다면 단순히 표 1의 내용을 그대로 문장으로 나열하고 설명하는 것 밖에 안되기 때문에, 세부사항 별 비율들은 간단히 표로 제공하고, 이 중 가장 많이 사용된 것과 흥미로웠던 것을 언급하고, 특히 흥미로웠던 것들에 대해서 이론적 설명을 덧붙이겠다.

Graphical user interface, text, application, email

Description automatically generated

표 1 : 세부사항 별 사용방법론 비율

**MODELLING**

Bayesian Joint Modelling를 다룬 논문에서 가장 많이 사용된 모델은 Linear Mixed Model이고, 흥미롭게 읽었던 모델은 Latent Variable Model이다 (각 비율은 표 2 참고).

Text, letter

Description automatically generated

표 2 : 모형 별 논문출현 비율

여기서 Latent Variable이란 관측되지 않았으나 Y변수에 영향을 줄 것이라 예상되는 변수를 뜻하며, Latent Variable Modelling의 주요 목적은 3가지가 있다. 첫 째는 기존 변수로 설명되지 않는 분산을 잡아내는 거이고, 둘 째는 예측 변수에 직접적으로 영향을 주는 변수를 만드는 것, 그리고 마지막은 예측 변수를 직접적으로 설명하는 변수를 만드는 것이다. Latent Variable Modelling의 대표적 예시로 Factor Analysis가 있다.

**RANDOM EFFECT DISTRIBUTION**

랜덤효과 분포 가정에 사용된 분포들의 비율은 표 3에 나와 있으며, 가장 많이 사용된 분포는 정규분포이고, 흥미로웠던 가정은 Dirichlet Process Prior이다.

Background pattern

Description automatically generated with low confidence

표 3 : 랜덤효과 분포 가정 비율

Dirichlet Process는 단순하게 말하자면 분포들의 분포로, Dirichlet Process를 이용해 sampling을 진행한다는 것은 분포를 sampling한다는 의미로 받아들이면 된다.

그림 2를 참고하며 설명하겠다. Sampling에서 사용하는 것은 Base Distribution 이 아니라, 과 같은 support를 갖는 G를 이용해서 sampling을 진행하는 것이다. 이 때, Dirichlet Process는 sample인 에 대한 분포가 아닌 G에 대한 분포이기 때문에 이 아니라, 수식 1과 같이 표기해야 함에 유의해야 한다.

A picture containing diagram

Description automatically generated

그림 2

Text

Description automatically generated

수식 1

이를 수식 2처럼 베이즈 정리를 이용해 정리하고, 에 대하여 marginalization을 시켜주면, 이 어떤 형태를 취하는지 알 수 있다 (결과는 수삭 3에 나와 있다). 그리고 이를 K개의 sample로 확장하면 수식 4처럼 나오는 것을 확인할 수 있다.

Text

Description automatically generated with low confidence

수식 2

Logo

Description automatically generated

수식 3

Text, letter

Description automatically generated

수식 4

**ERROR DISTRIBUTION**

오차의 분포에 대한 가정으로 가장 많이 사용된 것은 당연하게도 정규 분포이다. 48.6%로 압도적으로 높았으며, 이 중 흥미로웠던 것은 8.1%의 Skew Normal Distribution이었다. Skew Normal Distribution은 이상치가 많을수록 Robust한 결과를 출력하기 때문에 사용하고, 자세한 설명은 그림 3에 나와 있다. 이를 통해 알 수 있는 것은 정규분포는 Skew Normal Distribution의 한 부분이라는 것이다. ( = 0)

Graphical user interface

Description automatically generated with medium confidence

그림 3

**BAYESIAN SAMPLING ALGORITHM**

샘플링 기법으로 가장 많이 사용된건 38.8%로 MCMC이고(표 4), 이 중 눈에 띄었던 것은 1.4%의 점유율을 가졌던 Hamiltonian Monte Carlo였다. 이름에서 알 수 있듯이 MCMC의 변형으로, transition 생성방법과 적분을 근사하는 방식이 매우 다르다. Hamiltonian Monte Carlo는 대략 3단계로 구성되어 있다. 3단계를 설명하기 전에 알아야 할 개념인 Auxiliary Momentum Variable과 Hamiltonian에 대해서는 설명 1과 설명 2로 대체하겠다.

1단계는 는 current value = , 는 로initial value를 정하고, Hamiltonian 을 각각 와 에 대해 미분한다. 2단계는 미분을 통해 얻은 수식들을 사용하여 ) , 로 업데이트 하고, 이러한 과정을 L번 반복한 뒤, 로 정의한다. 그리고 마지막 3단계에서는 을 계산하여 1 이상이면 와 를 표본으로 받아들이고, 만약 1보다 작으면 와 를 initial value로 설정하고 1, 2단계를 반복한다.

Text

Description automatically generated with medium confidence

표 4

Text

Description automatically generated

설명 1

Text, letter

Description automatically generated with medium confidence

설명2

**CONCLUSION**

사실 처음 논문을 읽었을 때 예상했던 것보다 지나치게 이론적 그리고 수리적 내용이 없어서 당황하였다. 해당 논문은 단순히 저자 자신이 찾은 논문들을 읽고, 이에 대한 statistical summary만 제공했다고 해도 무방할 정도로 논문의 구조가 매우 단순했고, 사실 내용이랄게 없었다고 생각했다. 논문 그대로를 반영하여 발표를 하고, 이에 대한 보고서를 쓰기에는 무리라고 판단되어, 고민한 끝에 위와 같은 방식을 선택하게 되었다. 사실 아직 이론적으로 이해가 안가는 부분이 많고, 제대로 설명한 것인지 확신이 서지 않지만, 흥미로운 이론, 방법론, 분포에 대해 새로 알아냈고, 공부할 수 있었다는 사실에 만족한다.